Cross-Lingual Projection of LFG F-Structures: Building an F-Structure Bank for Polish

Alina Wróblewska† and Anette Frank‡

alina.wroblewska@ipipan.waw.pl frank@cl.uni-heidelberg.de

†Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

‡Department of Computational Linguistics, Heidelberg University, Germany

Abstract

Various methods aim at overcoming the shortage of NLP resources, especially for resource-poor languages. We present a cross-lingual projection account that aims at inducing an annotated treebank to be used for parser induction for Polish. Our approach builds on Hwa et al.'s projection method [7] that we adapt to the LFG framework. The goal of the experiment is the induction of an LFG f-structure bank for Polish. The projection yields competitive results. The resulting f-structure bank may be used to train a dependency parser for Polish, or for automatic induction of a probabilistic LFG grammar.

1 Introduction

Natural language processing has made rapid progress over the last decades. Yet, computational linguistic resources and tools are restricted to a handful of languages. The creation of high-quality resources for all languages using traditional, manual techniques is a time-consuming and expensive process. This holds especially for the creation of grammars and syntactic treebanks. Various methods aim at overcoming the shortage of NLP resources (such as bootstrapping, unsupervised learning, cross-lingual projection). The approach pursued in this paper targets the induction of linguistic annotations in a cross-lingual setting: Using a bilingual corpus, existing analysis tools are applied to the resource-rich language side of a bitext. The resulting annotations are projected to the second, resource-poor language via automatically produced word alignment links. This annotation projection approach for resource induction is built on the assumption that the linguistic analysis of a sentence carries over to its translation in an aligned parallel corpus. While this assumption does not hold uniformly, the projected annotations can be used to train NLP tools for the target language (cf. Section 2).

Within the ParGram project [3], grammars for English, French, German, Norwegian, Japanese, Urdu and further languages are written according to the framework of Lexical Functional Grammar (LFG), using the Xerox Linguistic Environment (XLE) as a processing platform. Manual development of large-scale LFG

grammars is an expensive process that may be sped up by automation techniques. One strand of work that targets the automatic induction of LFG grammars is the induction from existing syntactic treebanks (Cahill et al. [4]). However, this method relies on the availability of high-quality treebanks. To overcome the need of manual creation of treebanks, we investigate the cross-lingual projection approach to induce syntactically annotated corpora for new languages. Given the considerable divergence of constituent structures across languages, the grammar architecture of LFG, with its strong lexicon component and multiple levels of representations seems especially suited for a cross-lingual grammar induction task. F-structures are largely invariant across languages, and are thus especially suited to serve as the pivot for cross-lingual syntactic annotation projection. Following this insight, we pursue cross-lingual projection of grammatical functions to induce an f-structure bank for Polish. Our approach builds on Hwa et al. [7] and adapts their method to the LFG framework. We project the f-structure analyses of automatically parsed English sentences in a bitext to their Polish translations via word-alignment links. As the projected annotations are typically noisy, we apply a number of post-correction rules and filtering methods. The induced f-structure bank can be used to train a dependency parser for Polish. A full-fledged LFG grammar for Polish may be obtained by mapping the induced f-structures to appropriate c-structures (cf. Klein [8]), and using the obtained c- and f-structure bank for automatic LFG grammar induction, following the method of Cahill et al. [4].

Our paper is structured as follows. Section 2 gives an overview of the state of the art relevant to this work: We introduce the theoretical assumptions and related works on cross-lingual projection. We then focus on the projection of syntactic dependency relations and review existing computational linguistic resources for Polish. In Section 3 we outline relevant facts about the Polish language. We present the transposition of Hwa et al.'s work to the LFG framework and describe its application to the Polish language. Section 4 presents the data and experiments we conducted to induce an f-structure bank for Polish. Finally, we carry out some error analysis and compare our results to related work. Section 5 concludes.

2 State of the Art

2.1 Cross-lingual Annotation Projection

The cross-lingual annotation projection method consists in applying available monolingual NLP resources and tools in a multilingual scenario. Existing analysis tools are applied to the source language side of a word-aligned parallel corpus. Based on the assumption that the linguistic analysis of a sentence carries over to its translation in the bitext, the resulting linguistic annotations are projected from the source language onto the target language using automatic word alignment as a bridge. Annotation projection results in an automatically annotated corpus in the target language that can be used for supervised induction of NLP tools.

While the underlying assumption of cross-lingual correspondence is rather

strong, the cross-lingual projection method has been successfully applied to various levels of linguistic analysis and corresponding NLP tasks, such as PoS tagging and NP bracketing (Yarowsky and Ngai [14]), syntactic dependency annotation and parser induction (Hwa et al. [7], Ozdowska [10]), argument identification (Bouma et al. [1]), word sense disambiguation (Diab and Resnik [6]), semantic role labelling (Padó and Lapata [11]) and temporal labeling (Spreyer and Frank [12]).

2.2 Projection of Syntactic Dependencies

As shown in the pioneering work of Hwa et al. [7], syntactic dependencies are especially suitable for cross-lingual projection of syntactic information, as dependency relations can carry information across languages with varying word order. Specifically, the projection of syntactic dependencies is based on the *Direct Correspondence Assumption*, which states that the dependencies in a source sentence directly map to the syntactic relationships in the word-aligned target translation:

> Given a pair of sentences E and F that are (literal) translations of each other with syntactic structures $Tree_E$ and $Tree_F$, if nodes x_E and y_E of $Tree_E$ are aligned with nodes x_F and y_F of $Tree_F$, respectively, and if syntactic relationship $R(x_E, y_E)$ holds in $Tree_E$, then $R(x_F, y_F)$ holds in $Tree_F$. Hwa et al. (2005:314) [7]

Hwa et al. [7] apply this method for annotation projection and the induction of dependency parsers for Spanish and Chinese. In their experiments, the English side of a word-aligned parallel corpus is annotated with dependency representations. These annotations are directly projected onto the target language side. Since the direct projections are noisy, the projected representations are post-processed using about 12 language-specific correction rules. The transformed representations are used to train dependency parsers for Spanish and Chinese. The results obtained by Hwa et al. [7] will be presented in Section 4, in comparison to our results.

Two further studies apply cross-lingual methods for the induction of dependency relations. Ozdowska [10] projects part-of-speech tags, morphological information (gender and number) and syntactic dependencies from two source languages (English or French) onto one target language (Polish) using only one-to-one word alignments. The aim of this experiment is to verify which source language (English or French) is more suitable for the projection of particular annotations onto Polish. Even though the annotations for both source languages are available, Ozdowska does not test whether projection from both languages in a triangulation architecture could increase the results. By contrast, the triangulation method is explored in the multi-parallel annotation projection architecture proposed by Bouma et al. [1]. Here, selected grammatical functions are projected from (one or several) source language(s) (German, English) with the aim of verb argument identification in the target language (Dutch). The approach is similar to ours, since it is based on the Pargram LFG grammars and the target of the projection are grammatical

functions. However, the goal of our experiment is the induction of full-fledged f-structures, as opposed to verb-argument functions as in Bouma et al. [1]

In our work, we build on Hwa et al.'s approach that combines direct projection with language-specific post-correction rules. These rules target principled differences between the source and target languages that the DCA fails to capture, and thus offer a focused way for improving precision without impeding recall. Our account may still be complemented by triangulation techniques in later stages.

2.3 Computational Linguistic Resources for Polish

Polish is a language with relatively few NLP resources and tools. Currently, the following resources are available for Polish: corpora (e.g. the morpho-syntactically annotated corpus IPI PAN, PWN Corpus), morphological analyzers (e.g. Morfeusz, SAM, Morfologik), stemmer (Lametyzator), parsers (a DCG parser Świgra, a shallow parsing and disambiguation system Spejd)[1] and the Polish Wordnet Słowosieć.[2] Thus, there is a strong need for investigating methods for the rapid creation of further high-quality NLP resources for Polish.

3 Cross-lingual Induction of a Polish F-Structure Bank

3.1 Some Facts about Polish

In contrast to our source language English, Polish is rather unfamiliar to most readers and has been discussed and processed in few NLP studies. We briefly outline the main characteristics of the Polish syntax, with focus on syntactic phenomena that may be relevant to the projection task.

Morphology and word order. Polish is an inflecting language with relatively free word order and morphological identification of grammatical functions, by assignment of case. Thus, constituent order is rather flexible, as seen in (1.a,b).

(1) a. *Tomek kocha Marię.*
Tom.NOM-SUBJ love.3.SG Mary.ACC-OBJ
'Tom loves Mary.'

b. *Marię kocha Tomek.*
Mary.ACC-OBJ love.3.SG Tom.NOM-SUBJ
'Tom loves Mary.'

Pro-drop. As a pro-drop language, Polish allows omission of a personal pronoun in SUBJ function. The morphological features of the omitted subject are specified by the verb; the SUBJ is represented by an 'empty' pronoun PRED = 'pro', see (2).

[1] See overview on ACL Wiki: http://aclweb.org/aclwiki/index.php?title= Resources_for_Polish.
[2] See plWordNet Słowosieć at http://www.plwordnet.pwr.wroc.pl/browser/?lang=en

(2) *Gotuję* *obiad*.
cook.1.SG.PRES dinner.MASK.SG.ACC
'I am cooking a dinner.'

$$\begin{bmatrix} \text{PRED} & \text{'gotować<SUBJ,OBJ>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'obiad'} \\ \text{CASE} & \text{ACC} \end{bmatrix} \end{bmatrix}$$

Null specifiers. Polish does not possess articles corresponding to 'a/an' and 'the' in English that function as SPECifiers. However, there are some pronouns that may function as determiners in SPECifier function: possessive (*mój* 'my', *twój* 'your', etc.), demonstrative (*ten* 'this.MASK', *ta* 'this.FEM', etc.), quantificational (*niektórzy* 'some.MASK', *wielu* 'many.MASK', etc.), interrogative (*jaki* 'what.MASK', *która* 'which.FEM', etc.).

Negation and case marking. **(a)** In the so-called **genitive of negation**, an accusative-marked OBJect changes to genitive case if the verb is negated, while other arguments of the verb remain unaltered. In (3.a) the verb *czytać* (engl. 'to read') requires an accusative OBJect argument, while under negation in (3.b), the only possible case of the OBJ argument is genitive.

(3) a. *Czytam* *ksiażkę*.
read.1.SG.PRES book.ACC
'I'm reading a book.'

b. *Nie czytam* *książki*.
not read.1.SG.PRES book.GEN
'I'm not reading any book.'

(b) A special phenomenon of case marking called **feature indeterminacy** (Dalrymple and Kaplan [5]) is observed in coordination constructions such as (4).

(4) *Kogo* *Jan* *lubi* *a* *Jerzy* *nienawidzi*?
who.ACC/GEN John.NOM likes.3.SG.PRES Cnj George.NOM hates.3.SG.PRES
'Who does Jan like and George hate?'

In this type of construction, two verbs with conflicting object case marking are coordinated: *lubić* (engl. 'like') requires accusative object marking, while *nienawidzić* (engl. 'hate') requires its object to be marked genitive. The interrogative pronoun *kogo* (engl. 'who') fulfills the OBJ function that is distributed over the coordinated verb predicates. It can do so because of case syncretism, i.e. because it shows the same surface form both in accusative and genitive case. Dalrymple and Kaplan [5] account for this phenomenon in a set-based theory of case-marking, by defining indeterminate (or a set of) case(s) for *kogo*, as seen in (5).

	kogo	'who':	$(\uparrow \text{CASE}) = \{\text{ACC}, \text{GEN}\}$
(5)	*lubić*	'like$\langle$SUBJ, OBJ$\rangle$':	$\text{ACC} \in (\uparrow \text{OBJ CASE})$
	nienawidzić	'hate$\langle$SUBJ, OBJ$\rangle$':	$\text{GEN} \in (\uparrow \text{OBJ CASE})$

Since our account will be based on the projection of grammatical functions, genitive of negation or feature indeterminacy as special cases of case marking will not constitute any problem, as the grammatical functions involved remain constant.[3] Missing subjects and specifiers constitute structural divergences at the surface level and need to be accounted for in the projection module that considers special word alignment configurations (see below, Section 3.2). The fact that Polish has a relatively free word order fits nicely with LFG's conception of f-structures and grammatical functions, which are represented independently of surface word order.

3.2 Cross-lingual Projection of LFG F-Structures

The aim of our work is to create an LFG f-structure bank for Polish with minimal human involvement. We build on Hwa et al.'s approach and adapt their method of projecting dependency structures to the LFG framework. Two main characteristics of LFG make it especially suitable for this cross-lingual projection method: (i) Since LFG is a lexicalized theory, projection of annotations assigned to particular words can be sufficently guided by word alignment. (ii) F-structures constitute an abstract level of analysis that is largely invariant across languages, and thus perfectly suited for projection between languages with varying word order.

LFG f-structures encode grammatical functions holding between PREDicates and their arguments or modifiers, represented as partial f-structures. Next to PREDs, partial f-structures encode morpho-syntactic information, such as CASE, TENSE, PERSon and NUMber. The close correspondence between grammatical function information in LFG and syntactic dependencies in general is most easily seen by viewing grammatical functions holding between partial f-structures as relating two lexical items (PREDs) that head the corresponding partial f-structures.[4]

Cross-lingual projection of f-structures is defined as follows: Projection is grounded in automatically induced word alignment links $al(te_i, tf_i)$ ($al \in AL : E \times F$) between English and Polish surface words (terminals) $te_i \in E, tf_i \in F$. Based on the LFG-parsed English corpus, we identify the corresponding lexical items (PREDs) e in the English f-structures. The set GF consists of grammatical functions SUBJ, OBJ, OBL, ADJ, etc. that hold between pairs of English PREDicates $gf(e_i, e_j)$. During projection, grammatical functions gf encoded in the source f-structure fs_e are transferred to the target sentence via word alignment links, according to the following definition, which is similar to the one in Hwa et al. [7]:

> The grammatical function $gf(e_i, e_j)$ holding between PREDs e_i and e_j in the source f-structure fs_e projects to the target f-structure fs_f as $gf(f_i, f_j)$ if and only if the source terminals te_i and te_j that project to the PRED values e_i, e_j included in fs_e are aligned with the target terminals tf_i and tf_j of f_i and f_j, respectively.[5]

[3] However, since Polish has an extended case system distinguishing seven cases, studying the interaction of grammatical functions and case marking will be important for further improvements.

[4] In general, dependency structures are assumed to be trees, whereas LFG f-structures are graphs. This difference has no effect on the present projection approach.

The definition states that if two English words are related by a grammatical function, the same grammatical function will relate their word counterparts in Polish. Similar to Hwa et al. [7], we define specific projection constraints for different types of alignment links:[6]

one-to-one: a grammatical function $gf(e_i, e_j)$ relates source words e_i and e_j that are aligned with exactly one target word $al(e_i, f_i)$ and $al(e_j, f_j)$, respectively.

one-to-many: a grammatical function $gf(e_i, e_j)$ relates source words e_i and e_j that are aligned with a single target word: $al(e_i, f_i)$ and $al(e_j, f_i)$. The gf is projected to the partial target f-structure fs_i headed by f_i: $gf(f_i, f_j)$, where f_j is defined as [PRED 'pro'] for an incorporated SUBJ pronoun (pro-drop) or as [PRED 'null'] for other grammatical functions (e.g. null specifiers SPEC).

unaligned e: a grammatical function $gf(e_i, e_j)$ relates source words e_i and e_j, where e_i is aligned with the target word f_i $al(e_i, f_i)$, the other with *none* $al(e_j$, *none*). The gf is projected to the partial f-structure fs_i headed by f_i: $gf(f_i, f_j)$. f_j is defined as [PRED 'pro'] for an incorporated SUBJ pronoun or as [PRED 'null'] for other grammatical functions (e.g. null specifiers SPEC).

However, the *Direct Correspondence Assumption* that underlies the annotation projection approach is an idealisation. Indeed, the projected grammatical functions may be incorrect, due to

(i) errors in the source annotations obtained from automatic LFG parsing;

(ii) poor accuracy of automatic word alignment;

(iii) true mismatches of functional structure between English and Polish.

These error sources radically impair the quality of the projected grammatical functions. However, these shortcomings can be overcome by applying correction rules similar to Hwa et al. [7] that locally transform the induced Polish f-structures. We have defined two post-projection correction rules that are motivated by general linguistic properties of the Polish language:[7]

Rule 1: The PRED value of the SPEC_DET function for the article 'the' or 'a/an' in English is replaced by 'null' in the Polish partial f-structure (cf. Figure 1).

Rule 2: The grammatical function borne by an *of*-prepositional phrase in English is realized by a genitive noun phrase in Polish (cf. Figure 2).

[5]This definition presupposes lemmatisation on the target side, to provide appropriate values f_i for the target f-structure's PRED features. In practice, given that we don't use any tagger or morphological analyzer to preprocess the Polish corpus, the target f-structure PRED values are instantiated with the aligned Polish surface words (together with the English lemma, for better readability, see below Figures 1-3). In the following, we will use f_i, f_j instead of tf_i,tf_j, to keep the definitions simpler.

[6]Hwa et al. 2005 distinguish 5 alignment type scenarios: one-to-one, unaligned (English), one-to-many, many-to-one, many-to-many. In our unidirectional alignment experiment (cf. Section 4), we only found the 3 alignment types defined above.

[7]Further correction rules may be formulated by taking into account morpho-syntactic information concerning case, number, tense etc. Currently, we do not consider morpho-syntactic features.

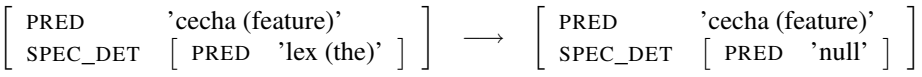

Figure 1: Example of application of Rule 1: an erroneously induced Polish article equivalent to English 'the' in SPEC_DET function is replaced by a 'null' specifier.

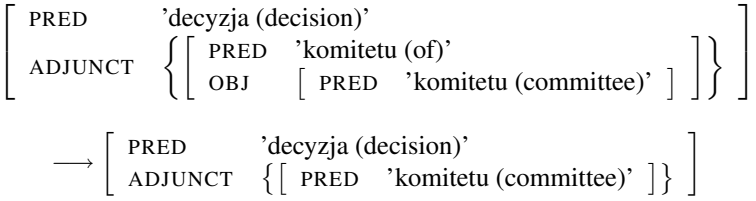

Figure 2: Example of application of Rule 2: the prepositional projection level of an *of*-PP in English is reduced to yield an NP adjunct in Polish.

We currently focus on the projection of grammatical functions, without considering morpho-syntactic features. The induced Polish f-structures are therefore *preds-only f-structures*. The following f-structure of the Polish sentence *Niniejsza dyrektywa skierowana jest do Państw Członkowskich* is automatically induced based on the f-structure of its English equivalent 'This directive is addressed to the Member States.'

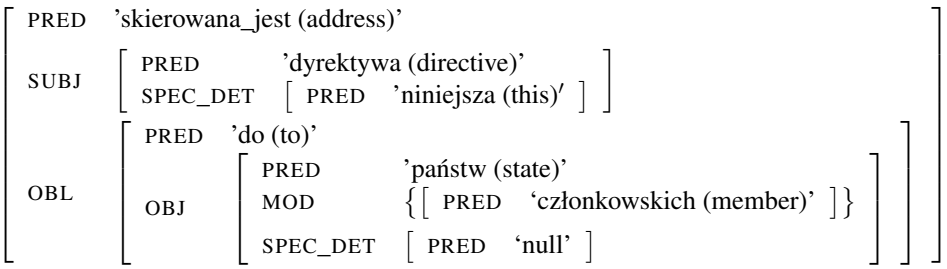

Figure 3: The f-structure of the Polish sentence *Niniejsza dyrektywa skierowana jest do Państw Członkowskich.* ('This directive is addressed to the Member States.')

Based on the projection of f-structure information as stated above, and enhanced with post-correction rules as seen in Figures 1 and 2, we obtain an LFG f-structure bank for Polish that may be used to train a dependency parser, or a full-fledged LFG c- and f-structure bank and LFG grammar, by inducing f- to c-structure mappings for Polish, along the lines of Klein [8] for English[8] and inducing a probabilistic treebank-based LFG grammar, using the method of Cahill et al. [4].

[8]The f-structures used in Klein [8] contain morpho-syntactic features based on PoS information. In a similar way, our proto-f-structures induced for Polish can be enriched with morphological information using a PoS-tagger for Polish (on the target language side).

4 Data, Evaluation and Results

4.1 Data and preprocessing

Our projection experiment is conducted on the JRC-Acquis Multilingual Parallel Corpus [13], a large collection of European Union legislative texts that – unlike Europarl – includes texts in Polish. From the full English-Polish section of JRC-Acquis (1,26 mil. sentence links) we selected a subcorpus aligned on the sentence level consisting of 257,144 sentence pairs. The average sentence length ranges from 4 to 30 tokens. In the preprocessing phase the parallel texts are word-aligned, the English side of the bi-text is parsed into LFG f-structures, and we apply some filtering methods. We briefly describe these phases, in turn.

Word alignment is performed with the SMT system MOSES [9], based on statistics captured from the entire corpus. To a certain degree English is an isolating language that makes use of function and non-content words. Polish, by contrast, is a highly inflecting language and needs in general fewer or as many words as English to express the same content. In order to decide whether alignment of one Polish word with one or more English words conforms to general translational mappings, we use unidirectional Polish-English word alignment as a basis for projection.

LFG parsing The English side of the parallel corpus is parsed with the hand-crafted wide-coverage English LFG grammar, which is enhanced with a statistical disambiguation component selecting the most probable analysis.

Filtering We filter all duplicates and omit sentences that contain inconsistencies of tokenization between MOSES and XLE.

4.2 Evaluation

We evaluate the quality of the automatically induced f-structures against a gold standard consisting of f-structures of 50 Polish sentences randomly selected from the entire corpus. The gold f-structures chosen from the preprocessed data (11.98 tokens/sent. for English, 9.76 tokens/sent. for Polish) have been manually corrected. For this purpose, the f-structures were first transformed to the SALSA/ TIGER XML format required by the SALTO annotation tool (Burchardt et al. [2]), which enables efficient modification by adding, deleting or correcting grammatical functions. We calculate precision, recall and f-score for various projection scenarios (cf. Table 1):

+/- correction: for exact match of projected grammatical functions, distinguishing direct projection (**direct**) and projection with post-modification (**+corr**) using the two correction rules mentioned above;

+/- automatic: for the projected grammatical functions taking into account automatically derived (noisy) word alignment (**automatic**) in contrast to hand-corrected (optimal) word alignment (**manual**), to establish an upper bound.

experiment	languages	sentences (corpus)	alignment	projection	LP /ULP	LR /ULR	F-score /UF-score
Current Experiment	en-pl	50 (JRC-Acquis)	automatic	direct	49/50	51/52	50/51
				+corr	64/64	63/63	63.5/63.5
			manual	direct	61/62	63/63	62/62.5
				+corr	85/85	81/82	83/83.5
Hwa et al.	en-sp	100 (UN/FBIS /Bible)	automatic	direct			/33.9
				+corr			/65.7
			manual	direct			/36.8
				+corr			/70.3
Ozdowska	en-pl	50 (AC)	automatic	direct	67/82		
Bouma et al.	ge-de	222 (Europarl)	automatic	direct	52.2	52.9	52.6
	en-du				54.3	48.8	51.4
	(ge,en)-du				74.6	34.1	46.8

Table 1: LP/ULP: precision for labeled/unlabeled grammatical functions; LR/ULR: recall for labeled/unlabeled grammatical functions.

Results. As expected, direct projection of grammatical functions is noisy (49.98% f-score). Application of language-specific transformation rules considerably improves the accuracy of the projected grammatical functions (63.5% f-score). The *quality of word alignment* is a crucial factor for projection quality: projection based on corrected word alignments enhances the quality of the induced f-structures by 12 percentage points (pp) f-score for direct projection and by 19.45 pp f-score for projection with transformation rules. In line with Hwa et al. [7], these results clearly indicate that direct projection of grammatical functions is significantly outperformed by projection using *post-projection transformation rules*, both for automatic word alignment (13.52 pp f-score improvement), and for manually corrected word alignments, the latter constituting an upper bound of 20.97 pp f-score improvement. The upper bound (projection based on perfect word alignment) in conjunction with two correction rules yields an accuracy of 82.95% f-score.

4.3 Error Analysis

According to our error analysis, most errors are due to word alignment mistakes, especially wrong alignment of a post-modifier of a Polish noun with an English head noun. Further, errors in the automatically parsed English f-structures have a big impact, as grammatical functions are projected to Polish without quality tests. A final source of errors are translational divergences. They are caused by language-specific conventions or structural constraints on how to express the same content in different languages, or simply by some "translational freedom" taken by translators. Translational divergences may radically change the syntactic structure of a translated sentence as compared to its source. Erroneous f-structures caused by word alignment errors, mistakes in the source analysis or translational divergences may be filtered on the target side, if morphological or PoS information is provided.

4.4 Comparison to Related Work

A full comparison to related approaches is difficult, due to the different languages and corpora involved. Keeping this in mind, compared to the results of unlabeled dependency projection for Spanish in Hwa et al. [7], we obtain comparable f-scores (63.5 vs. 65.7) for automatic projection with post-correction. For direct projection, we outperform Hwa et al.'s results by 17 pp. Compared to Ozdowska [10], precision of our direct projection is lower by 18/32 pp. Ozdowska [10] relies on one-to-one alignment links (intersection) only, which increases precision but decreases recall. As Ozdowska [10] does not report recall figures, we cannot compare the results. Bouma et al. [1] represents an LFG-based approach, like ours. However, their work is restricted to verb arguments while we perform full f-structure induction. We observe that combining argument information from two languages (English and German) enhances precision but degrades recall. In contrast, we obtain balanced precision and recall values. Regarding f-score, our projection of grammatical functions outperforms the projection by Bouma et al. [1] by 16.7 pp.

5 Conclusions and Future Work

In summary, we presented a cross-lingual projection approach for creating an LFG f-structure bank for Polish. Our results are competitive as compared to related prior work on different languages and corpora. Similar to Hwa et al.'s work, the application of post-correction rules significantly improves the quality of the induced f-structures obtained by direct projection. It is worth mentioning that we obtain high, balanced precision and recall values. Based on the gold standard word alignment, we identified an upper bound of 83% f-score to build a Polish f-structure bank. We find that word alignment is a crucial factor affecting the accuracy of the projected grammatical functions, next to principled linguistic differences. In fact, there is a stark contrast between Hwa et al.'s delta to achieve the gold standard level (70.3% vs. 65.7%) and ours (83.5% vs. 63.5%). This indicates that word alignment constitutes a harder problem for alignment of English with Polish as compared to Spanish. Linguistic differences may be addressed by extending the set of post-correction rules. Word alignment may be improved by advances in the state of the art, e.g. using lemmatised and PoS-tagged corpora for word alignment.

In future work, we will explore inclusion of problematic data (inconsistent tokenization), improvement of word alignment, and use of morpho-syntactic information to further enhance the projection quality. The resulting Polish f-structure bank may be efficiently used to train a dependency parser. Training on Hwa et al.'s Spanish data (with 65.7% f-score) yielded a parsing performance of 72.1% f-score [7]. Since our projection model approaches comparable f-score (63.5%), we expect that a dependency parser for Polish will achieve comparable performance, with potential increase by further improvement of the base projection quality. Finally, we will explore induction of a full-fledged LFG grammar for Polish, by adding a module that learns f-to-c-structure mappings for Polish, along the lines of Klein [8].

References

[1] G. Bouma, J. Kuhn, B. Schrader, and K. Spreyer. Parallel LFG Grammars on Parallel Corpora: A base for practical triangulation. In M. Butt and T.-H. King, editors, *Proceedings of the LFG 2008 Conference*, pages 169–189, Sydney, 2008.

[2] A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. SALTO - A Versatile Multi-Level Annotation Tool. In *Proceedings of LREC 2006*, pages 517–520, 2006.

[3] M. Butt, H. Dyvik, T.H. King, H. Masuichi, and Rohrer Ch. The Parallel Grammar Project. In *Proceedings of the COLING 2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7, Taipei, Taiwan, 2002.

[4] A. Cahill, M. Forst, M. Burke, M. McCarthy, R. O'Donovan, C. Rohrer, J. van Genabith, and A. Way. Treebank-Based Acquisition of Multilingual Unification Grammar Resources. *Journal of Research on Language and Computation*, 3(2):247–279, 2005.

[5] M. Darlymple and R.M. Kaplan. Feature Indeterminacy and Feature Resolution. *Language*, 76(4):759–798, 2000.

[6] M. Diab and P. Resnik. An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, ACL 2002*, pages 255–262, Philadelphia, 2002.

[7] R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering*, 11(3):311–325, 2005.

[8] A. Klein. Von Dependenzstrukturen zu Konstituentenstrukturen: Automatische Generierung von Penn-Treebank-Bäumen aus LFG-F-Strukturen. Master's thesis, Universität Heidelberg, 2009.

[9] P. Koehn, M. Federico, W. Shen, N. Bertoldi, O. Bojar, Ch. Callison-Burch, B. Cowan, Ch. Dyer, H. Hoang, R. Zens, A. Constantin, Ch.C. Moran, and E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007*, pages 177–180, Prague, 2007.

[10] S. Ozdowska. Projecting POS Tags and Syntactic Dependencies from English and French to Polish in Aligned Corpora. In *Proceedings of the EACL 2006 Workshop on Cross-Language Knowledge Induction*, pages 53–60, Trento, 2006.

[11] S. Padó and M. Lapata. Cross-linguistic Projection of Role-semantic Information. In *Proceedings of HLT/EMNLP 2005*, pages 859–866, Vancouver, 2005.

[12] K. Spreyer and A. Frank. Projection-based Acquisition of a Temporal Labeller. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 489–496, Hyderabad, India.

[13] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, Tufis D., and D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC 2006*, pages 2142–2147, Genoa, Italy, 2006.

[14] D. Yarowsky and G. Ngai. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In *Proceedings of NAACL 2001*, pages 200–207, 2001.